Applicant Lincoln et al.
Title: Database for Storage and
Analysis of Full-Length Sequences
Filing Date: September 19, 2001
Serial No.: To Be Assigned
Atty Docket No: 06514069CON

FIG. 1

RNA Isolation Protocol

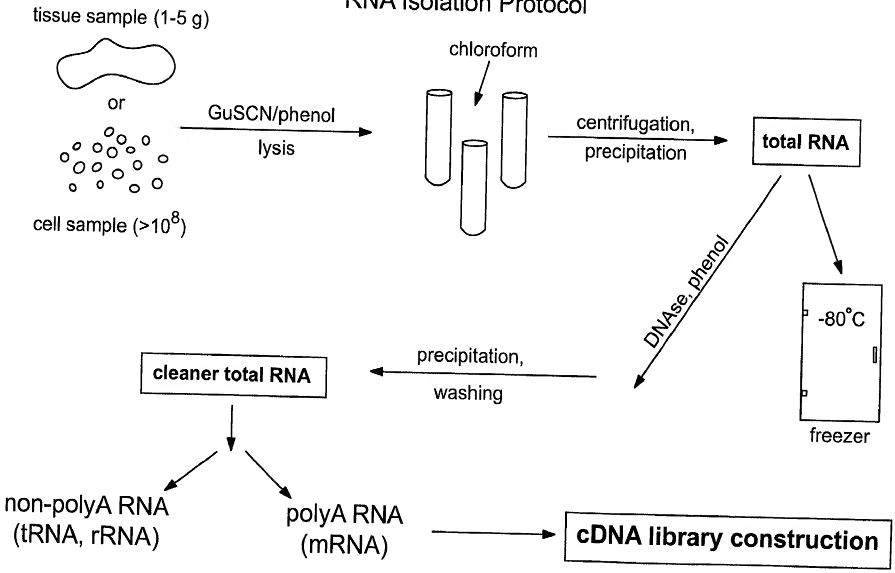
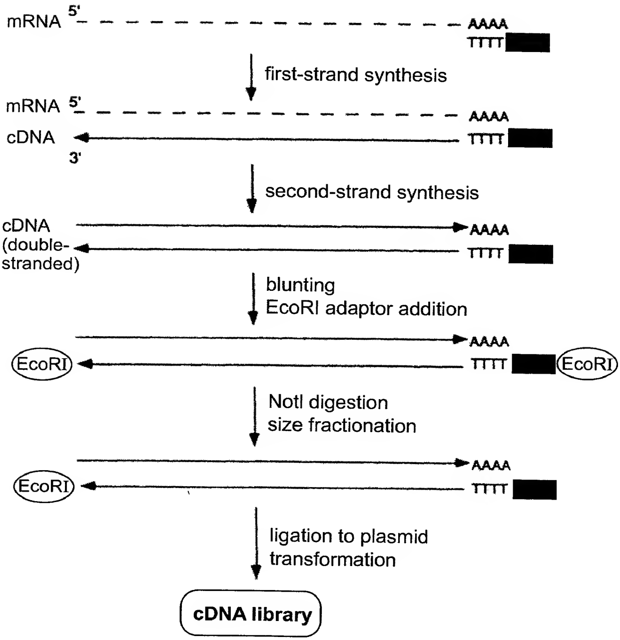


FIG. 2

cDNA Library Construction



Block 1 Sequence Editing Screens

Target Sequence Feature	Editing Method	Result
5' and 3' Vector	Dynamic Programming	Clip
PolyA Tail	Regular Expression	Clip
Sequencing Artifacts	Nearest Neighbor	Remove
Low Information	BLAST (S ≥ 90)	Mask
Contamination	BLAST (S ≥ 90)	Remove
Repetitive Elements	BLAST (S ≥ 90)	Mask
Mitochondrial	BLAST (S ≥ 90)	Remove
Ribosomal RNA	BLAST (S ≥ 90)	Remove

FIG. 3

FIG. 4

Aberrant

	A	C	G	T	
A	32	8	8	8	56
C	8	32	8	8	56
G	8	8	32	8	56
T	8	8	8	32	56
	56	56	56	56	224

Expected

	A	C	G	T	
A	14	14	14	14	56
C	14	14	14	14	56
G	14	14	14	14	56
T	14	14	14	14	56
	56	56	56	56	224

e.g. CCCCGGGGTTTTCCCCAAAAGGGG...

FIG. 5

Rigorous Statistics:

$$E = KNe^{-\lambda S} \text{ where } S = \sum_{HSP} s(a_i, b_j)$$

$s(a,a) = 5$ and $s(a,b) = -4$ for $a \neq b$

FIG. 6

Query	Match	BLAST Score	% ID	Length	Product Score
1969935 (238)	1798270	1103	99	259	91
	2057694	765	100	234	65
	1961765	600	100	245	50
	473298	318	92	261	25
473298 (261)	2057694	634	93	234	50
	1961765	634	93	254	46
	1969935	318	92	238	25
	1798270	254	94	259	18

Alignments

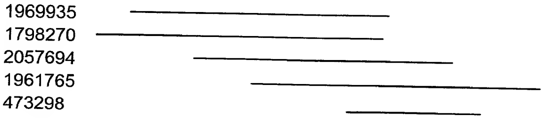


FIG. 7

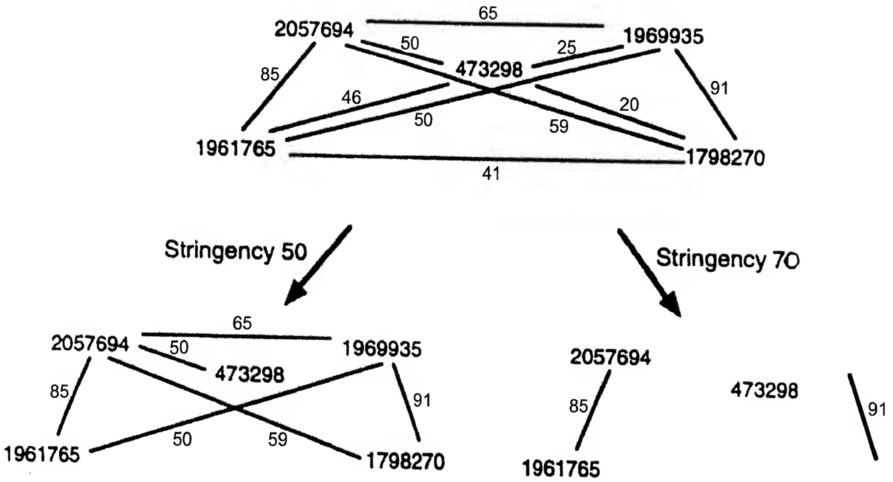

65
2057694
1969935
50
25
473298
85
46
20
91
50
59
1961765
1798270
41
Stringency 50
Stringency 70
65
2057694
50
1969935
473298
85
91
50
59
1961765
1798270
2057694
473298
85
91
1961765

FIG. 8

Creating a Master Cluster

Example: Cluster - 12 (singleton), Cluster 2, and Cluster 1 all contain representative clones with PS ≥ 40 to Gene X.

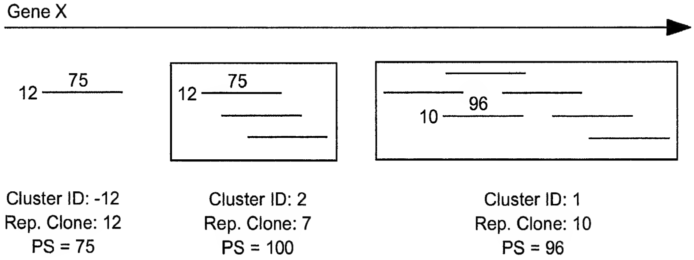
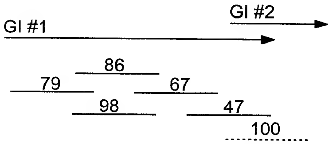


FIG. 9

Naming a Cluster

Q: Does any member of the cluster have a match to GenBank?

YES

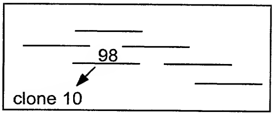


5 clones (solid line) match GI #1
1 clone (dashed) matches GI #2

Q: Which GI is represented the most
in the cluster?

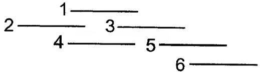
A: GI #1 is represented the most.

Q: Of those clones matching GI #1,
which has the highest Product Score?

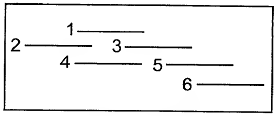


A: Clone 10 has a Product Score of 98,
so the cluster is named after this
representative clone.

NO



Q: Which clone has the lowest Clone
ID for this cluster?



unique cluster

A: Clone 1 has the lowest ID, so the
unique cluster is named after this
representative clone.

FIG. 10

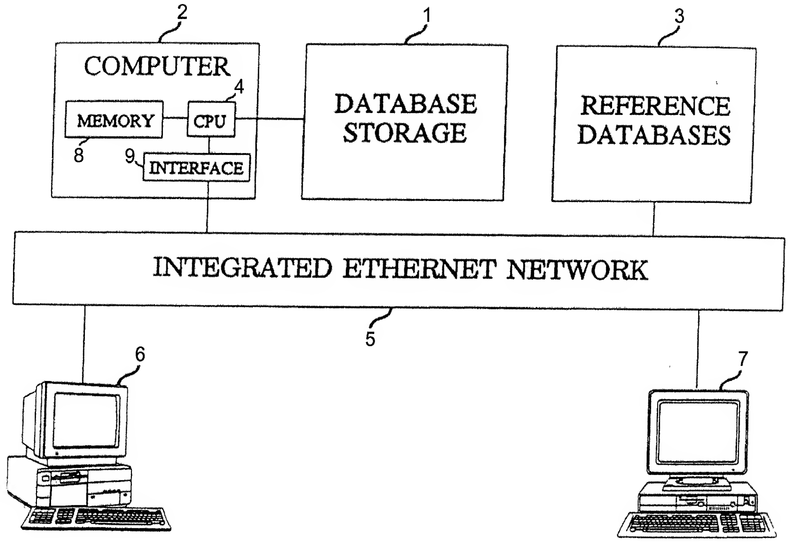

2
COMPUTER
4
MEMORY
CPU
8
9
INTERFACE
1
DATABASE STORAGE
3
REFERENCE DATABASES
INTEGRATED ETHERNET NETWORK
5
6
7

FIG. 11a

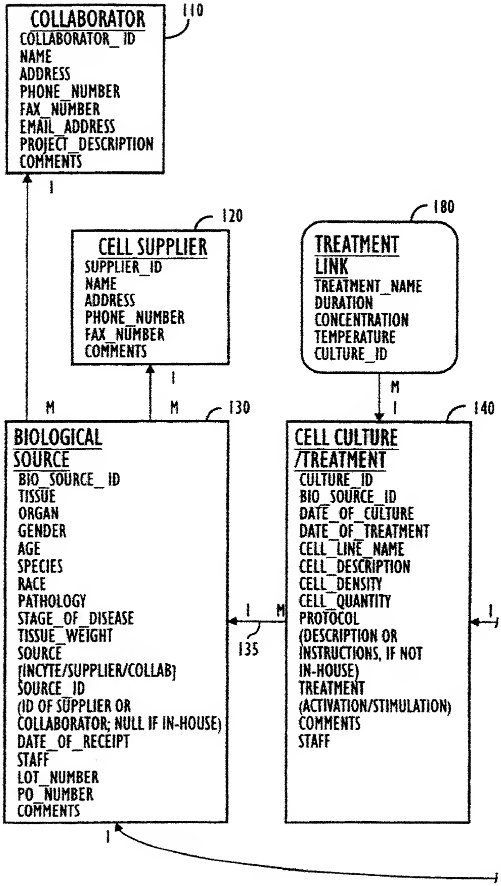
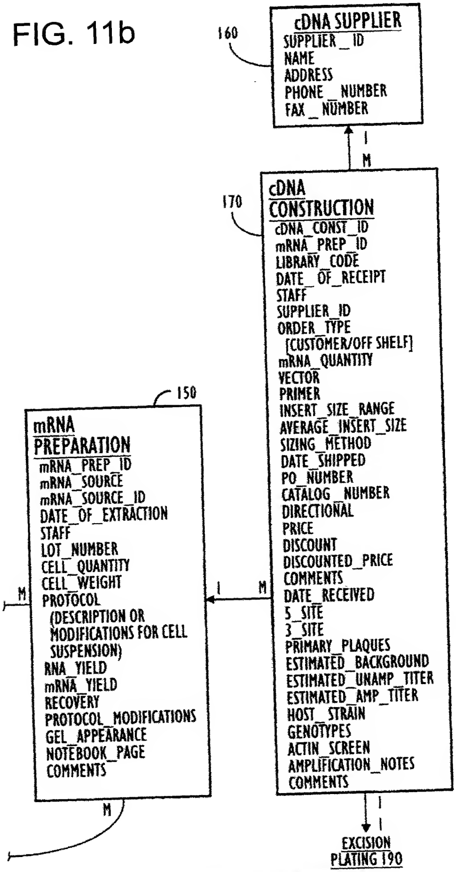
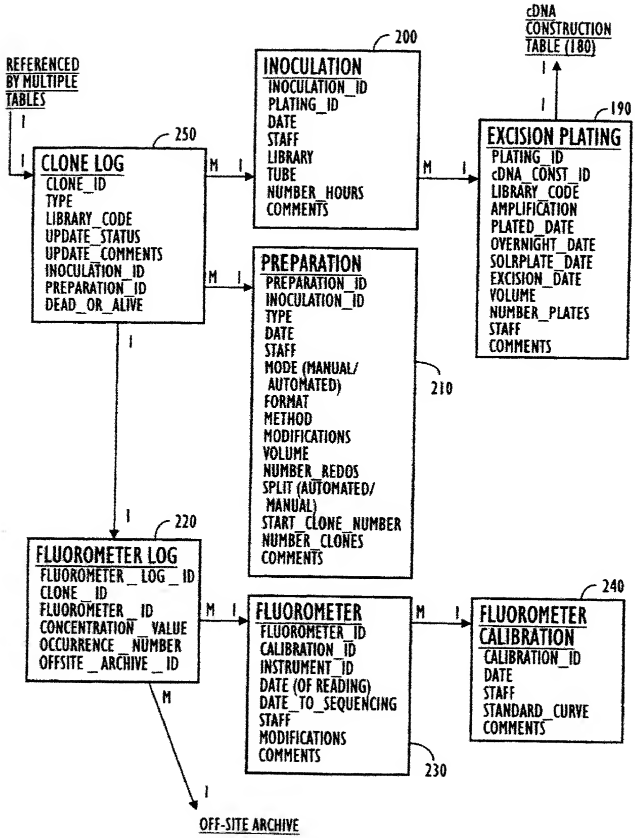

COLLABORATOR
COLLABORATOR_ID
NAME
ADDRESS
PHONE_NUMBER
FAX_NUMBER
EMAIL_ADDRESS
PROJECT_DESCRIPTION
COMMENTS
110
1
M
CELL SUPPLIER
SUPPLIER_ID
NAME
ADDRESS
PHONE_NUMBER
FAX_NUMBER
COMMENTS
120
1
M
TREATMENT LINK
TREATMENT_NAME
DURATION
CONCENTRATION
TEMPERATURE
CULTURE_ID
180
M
1
130
140
BIOLOGICAL SOURCE
BIO_SOURCE_ID
TISSUE
ORGAN
GENDER
AGE
SPECIES
RACE
PATHOLOGY
STAGE_OF_DISEASE
TISSUE_WEIGHT
SOURCE
[INCYTE/SUPPLIER/COLLAB]
SOURCE_ID
(ID OF SUPPLIER OR
COLLABORATOR; NULL IF IN-HOUSE)
DATE_OF_RECEIPT
STAFF
LOT_NUMBER
PO_NUMBER
COMMENTS
1
M
135
CELL CULTURE /TREATMENT
CULTURE_ID
BIO_SOURCE_ID
DATE_OF_CULTURE
DATE_OF_TREATMENT
CELL_LINE_NAME
CELL_DESCRIPTION
CELL_DENSITY
CELL_QUANTITY
PROTOCOL
(DESCRIPTION OR
INSTRUCTIONS, IF NOT
IN-HOUSE)
TREATMENT
(ACTIVATION/STIMULATION)
COMMENTS
STAFF
1
1

FIG. 11b


160
cDNA SUPPLIER
SUPPLIER_ID
NAME
ADDRESS
PHONE_NUMBER
FAX_NUMBER
1
M
170
cDNA CONSTRUCTION
cDNA_CONST_ID
mRNA_PREP_ID
LIBRARY_CODE
DATE_OF_RECEIPT
STAFF
SUPPLIER_ID
ORDER_TYPE
[CUSTOMER/OFF SHELF]
mRNA_QUANTITY
VECTOR
PRIMER
INSERT_SIZE_RANGE
AVERAGE_INSERT_SIZE
SIZING_METHOD
DATE_SHIPPED
PO_NUMBER
CATALOG_NUMBER
DIRECTIONAL
PRICE
DISCOUNT
DISCOUNTED_PRICE
COMMENTS
DATE_RECEIVED
5_SITE
3_SITE
PRIMARY_PLAQUES
ESTIMATED_BACKGROUND
ESTIMATED_UNAMP_TITER
ESTIMATED_AMP_TITER
HOST_STRAIN
GENOTYPES
ACTIN_SCREEN
AMPLIFICATION_NOTES
COMMENTS
150
mRNA PREPARATION
mRNA_PREP_ID
mRNA_SOURCE
mRNA_SOURCE_ID
DATE_OF_EXTRACTION
STAFF
LOT_NUMBER
CELL_QUANTITY
CELL_WEIGHT
PROTOCOL
(DESCRIPTION OR MODIFICATIONS FOR CELL SUSPENSION)
RNA_YIELD
mRNA_YIELD
RECOVERY
PROTOCOL_MODIFICATIONS
GEL_APPEARANCE
NOTEBOOK_PAGE
COMMENTS
M
1
M
M
1
1
EXCISION PLATING 190

Applicant Lincoln et al.
Title: Database for Storage and
Analysis of Full-Length Sequences
Filing Date: September 19, 2001
Serial No.: To Be Assigned
Atty Docket No: 06514069CON

FIG. 12



Applicant Lincoln et al.
Title: Database for Storage and
Analysis of Full-Length Sequences
Filing Date: September 19, 2001
Serial No.: To Be Assigned
Atty Docket No: 06514069CON

FIG. 13

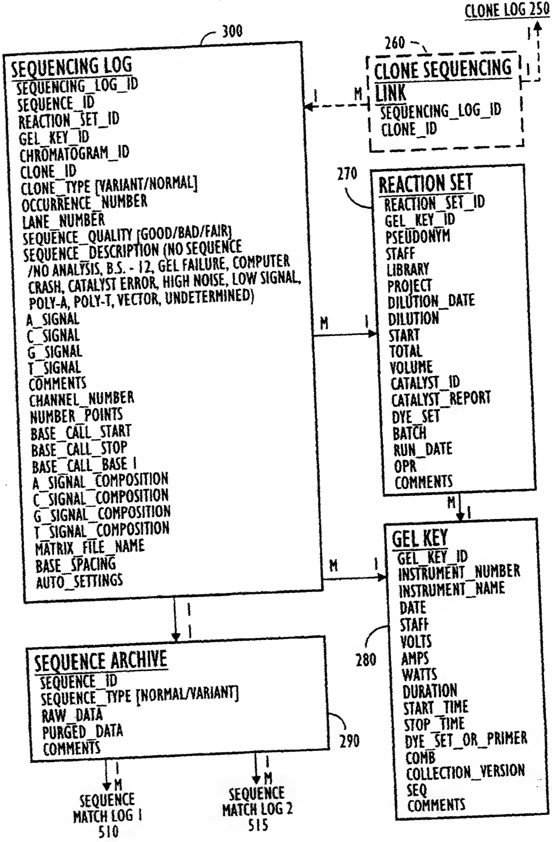


FIG. 14

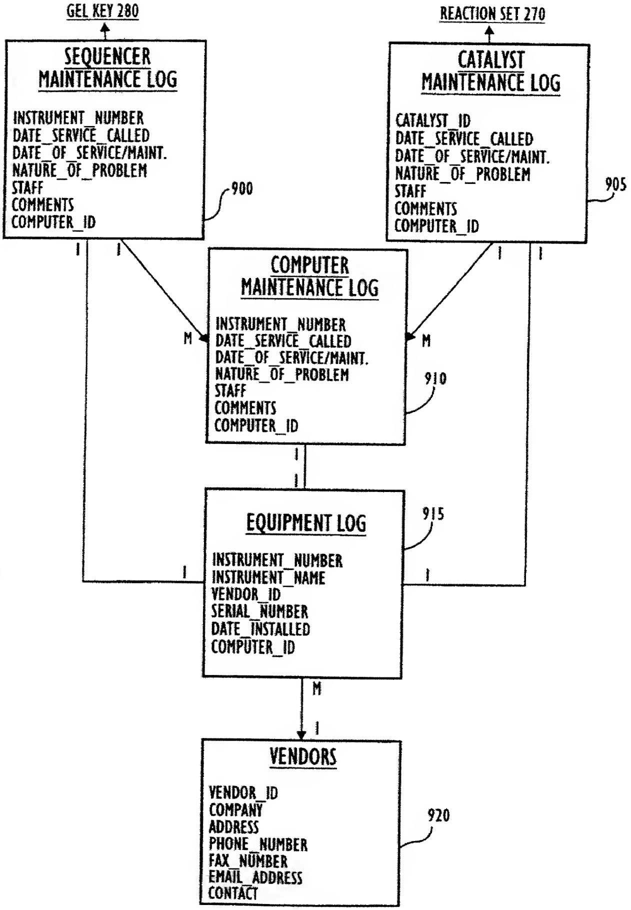
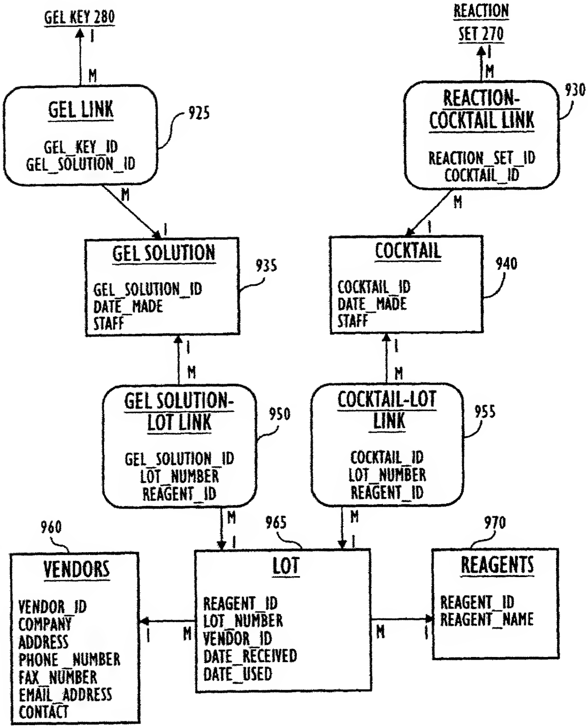

GEL KEY 280
REACTION SET 270
SEQUENCER MAINTENANCE LOG
INSTRUMENT_NUMBER
DATE_SERVICE_CALLED
DATE_OF_SERVICE/MAINT.
NATURE_OF_PROBLEM
STAFF
COMMENTS
COMPUTER_ID
900
CATALYST MAINTENANCE LOG
CATALYST_ID
DATE_SERVICE_CALLED
DATE_OF_SERVICE/MAINT.
NATURE_OF_PROBLEM
STAFF
COMMENTS
COMPUTER_ID
905
1
1
1
1
M
M
COMPUTER MAINTENANCE LOG
INSTRUMENT_NUMBER
DATE_SERVICE_CALLED
DATE_OF_SERVICE/MAINT.
NATURE_OF_PROBLEM
STAFF
COMMENTS
COMPUTER_ID
910
1
1
EQUIPMENT LOG
INSTRUMENT_NUMBER
INSTRUMENT_NAME
VENDOR_ID
SERIAL_NUMBER
DATE_INSTALLED
COMPUTER_ID
915
1
1
M
1
VENDORS
VENDOR_ID
COMPANY
ADDRESS
PHONE_NUMBER
FAX_NUMBER
EMAIL_ADDRESS
CONTACT
920

FIG. 15


GEL KEY 280
1
M
GEL LINK
GEL_KEY_ID
GEL_SOLUTION_ID
925
M
1
REACTION SET 270
1
M
REACTION-COCKTAIL LINK
REACTION_SET_ID
COCKTAIL_ID
930
M
1
GEL SOLUTION
GEL_SOLUTION_ID
DATE_MADE
STAFF
935
COCKTAIL
COCKTAIL_ID
DATE_MADE
STAFF
940
1
M
GEL SOLUTION-LOT LINK
GEL_SOLUTION_ID
LOT_NUMBER
REAGENT_ID
950
1
M
COCKTAIL-LOT LINK
COCKTAIL_ID
LOT_NUMBER
REAGENT_ID
955
M
1
965
M
1
960
VENDORS
VENDOR_ID
COMPANY
ADDRESS
PHONE_NUMBER
FAX_NUMBER
EMAIL_ADDRESS
CONTACT
1
M
LOT
REAGENT_ID
LOT_NUMBER
VENDOR_ID
DATE_RECEIVED
DATE_USED
M
1
970
REAGENTS
REAGENT_ID
REAGENT_NAME

Applicant Lincoln et al.
Title: Database for Storage and Analysis of Full-Length Sequences
Filing Date: September 19, 2001
Serial No.: To Be Assigned
Atty Docket No: 06514069CON

FIG. 16

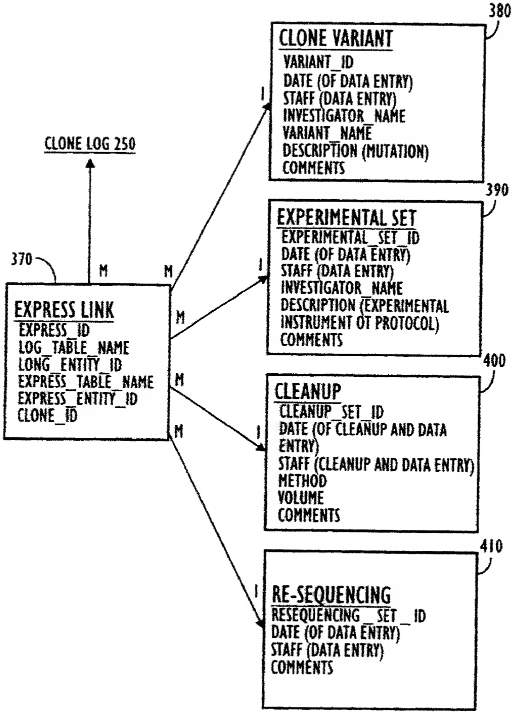


FIG. 17

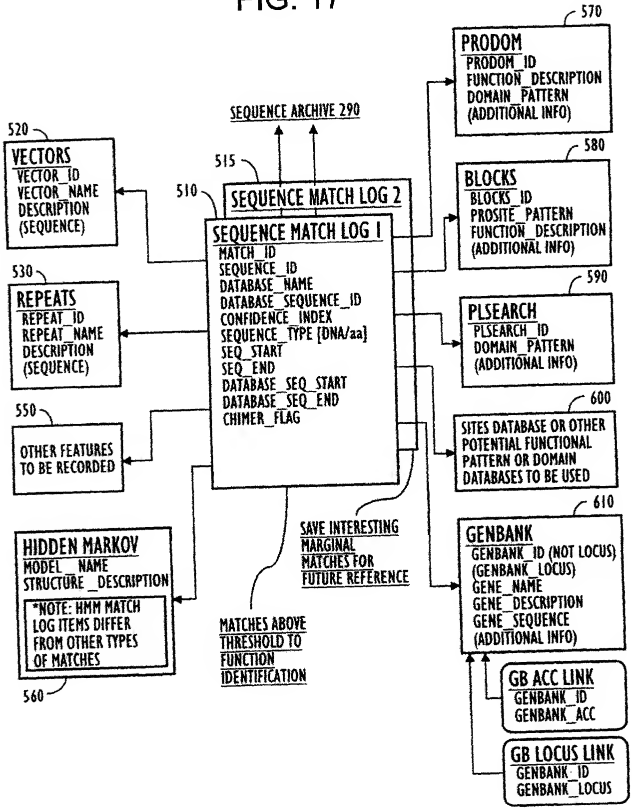
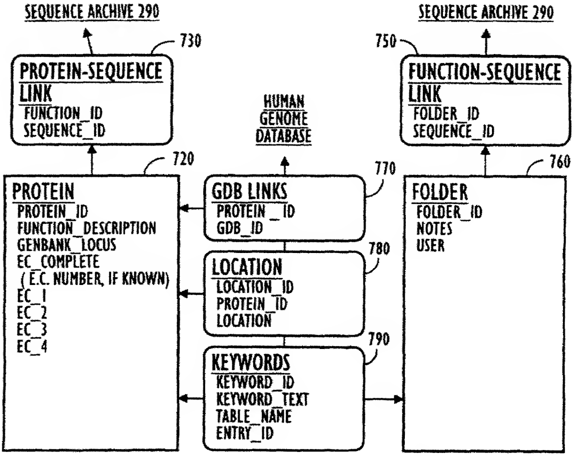

570
PRODOM
PRODOM_ID
FUNCTION_DESCRIPTION
DOMAIN_PATTERN
(ADDITIONAL INFO)
SEQUENCE ARCHIVE 290
520
VECTORS
VECTOR_ID
VECTOR_NAME
DESCRIPTION
(SEQUENCE)
515
510
SEQUENCE MATCH LOG 2
580
BLOCKS
BLOCKS_ID
PROSITE_PATTERN
FUNCTION_DESCRIPTION
(ADDITIONAL INFO)
SEQUENCE MATCH LOG 1
MATCH_ID
SEQUENCE_ID
DATABASE_NAME
DATABASE_SEQUENCE_ID
CONFIDENCE_INDEX
SEQUENCE_TYPE [DNA/aa]
SEQ_START
SEQ_END
DATABASE_SEQ_START
DATABASE_SEQ_END
CHIMER_FLAG
530
REPEATS
REPEAT_ID
REPEAT_NAME
DESCRIPTION
(SEQUENCE)
590
PLSEARCH
PLSEARCH_ID
DOMAIN_PATTERN
(ADDITIONAL INFO)
550
OTHER FEATURES TO BE RECORDED
600
SITES DATABASE OR OTHER POTENTIAL FUNCTIONAL PATTERN OR DOMAIN DATABASES TO BE USED
610
HIDDEN MARKOV
MODEL_NAME
STRUCTURE_DESCRIPTION
*NOTE: HMM MATCH LOG ITEMS DIFFER FROM OTHER TYPES OF MATCHES
SAVE INTERESTING MARGINAL MATCHES FOR FUTURE REFERENCE
GENBANK
GENBANK_ID (NOT LOCUS)
(GENBANK_LOCUS)
GENE_NAME
GENE_DESCRIPTION
GENE_SEQUENCE
(ADDITIONAL INFO)
MATCHES ABOVE THRESHOLD TO FUNCTION IDENTIFICATION
560
GB ACC LINK
GENBANK_ID
GENBANK_ACC
GB LOCUS LINK
GENBANK_ID
GENBANK_LOCUS

FIG. 18


SEQUENCE ARCHIVE 290
730
PROTEIN-SEQUENCE LINK
FUNCTION_ID
SEQUENCE_ID
SEQUENCE ARCHIVE 290
750
FUNCTION-SEQUENCE LINK
FOLDER_ID
SEQUENCE_ID
HUMAN GENOME DATABASE
720
PROTEIN
PROTEIN_ID
FUNCTION_DESCRIPTION
GENBANK_LOCUS
EC_COMPLETE
(E.C. NUMBER, IF KNOWN)
EC_1
EC_2
EC_3
EC_4
770
GDB LINKS
PROTEIN_ID
GDB_ID
780
LOCATION
LOCATION_ID
PROTEIN_ID
LOCATION
790
KEYWORDS
KEYWORD_ID
KEYWORD_TEXT
TABLE_NAME
ENTRY_ID
760
FOLDER
FOLDER_ID
NOTES
USER